GENE EXPRESSION PROFILES IN LIVER CANCER

INVENTORS

Darci Horne, Uwe Scherf and Joseph Vockley

RELATED APPLICATIONS

This application is related to U.S. Provisional Application 60/211,379, filed on June 14, 2000, and U.S. Provisional Application 60/237,054, filed October 2, 2000, which are herein incorporated by reference in their entirety.

BACKGROUND OF THE INVENTION

Primary hepatocellular carcinoma (HCC) is a widespread cancer throughout the world, especially prevalent where the incidence of chronic hepatitis B (HBV) and hepatitis C (HCV) viral infections are endemic (Groen, (1999) Semin. Oncol. Nurs. 15, 48-57; Idilman *et al.*, (1998) J. Viral. Hepat. 5, 110-117; Di Bisceglie *et al.*, (1998) Hepatol. 28, 1161-1165; Johnson, (1997) Hepatogastroenerology 44, 307-312; Sheu, (1997) J. Gastroeneterol. Hepatol. 12, S309-313). Hepatocellular carcinomas are very malignant tumors that generally offer a poor prognosis, dependent on the size of the tumor, the effect on normal liver functions, and the involvement of metastases. They are best treated by surgical resection, when the tumors are diagnosed at a stage where this is a viable possibility, but the recurrence rate for these cancers remains high (Johnson, (1997) Hepatogastroenterology 44, 307-312; Schafer & Sorrell, (1999) Lancet 353, 1253-1257; Groen, (1999) Semin. Oncol. Nurs. 15, 48-57; Sitzman, (1995) World. J. Surg. 19, 790-794; DiCarlo, (1995) Hepato-Gastroenterol. 42, 222-259; Tanaka *et al.*, (1996) Hepato-Gastroenterol. 43, 1172-1181; El-Assal *et al.*, (1997) Surgery 122, 571-577).

Numerous risk factors for the development of HCC have been identified: cirrhosis, HBV or HCV infection, being male, alcohol-related liver disease, exposure to aflatoxins, vinyl chloride and radioactive thorium dioxide, cigarette smoking, ingestion of inorganic arsenic, the use of oral contraceptives and anabolic steroids, iron accumulation, and various

inherited metabolic disorders (hemochromatosis, glycogen storage disease, porphyria, tyrosinemia, α-1-antitrypsin deficiency) (Di Bisceglie *et al.*, (1998) Hepatol. 28, 1161-1165; Chen *et al.*, (1997) J. Gastroenterol. Hepatol. 12, S294-308; Schafer & Sorrell (1999) Lancet 353, 1253-1257; Groen, (1999) Semin. Oncol. Nurs. 15, 48-57; Idilman *et al.*, (1998) J. Viral. Hepat. 5, 110-117; Johnson, (1997) Hepato-Gastroenterol. 44, 307-312).

In addition to liver tumors attributed to hepatocellular carcinoma, there are liver tumors that arise as metastases from primary tumors in other parts of the body. These tumors most often metastasize from the gastrointestinal organs, primarily the colon and rectum, but it is possible for metastatic liver cancers to occur from primary cancers throughout the body (Sitzman 1990, Groen 1999). These cancers can be treated using the routine therapies such as chemotherapy, radiotherapy, surgical resection, liver transplantation, chemoembolization, cryosurgery, or a combination of therapies (Sitzman, (1990) Hepatic Neoplasia, in Bayless (editor) Current Therapy in Gastroenterology and Liver Disease, Marcel Dekker; Groen, (1999) Semin. Oncol. Nurs. 15, 48-57).

The characterization of genes that are differentially expressed in tumorigenesis is an important step in identifying those that are intimately involved in the details of a cell's transformation from normal to cancerous. Studies examining the gene expression of metastatic liver tumors and hepatocellular carcinomas in comparison with a set of normal liver tissues would produce data identifying genes that are not expressed in normal livers but have been switched on in tumors, as well as genes that have been completely turned off in these tumors during the progression from a normal to a malignant state. Such studies would also lead to the identification of genes that are expressed in tumor tissue at differing levels, but not expressed at any level in normal liver tissue. The identification of genes and ESTs that are expressed in both types of tumors, *i. e.*, primary hepatocellular carcinomas as well as metastatic tumors of a different origin, and not in normal liver cells would be extremely valuable for the diagnosis of liver cancer.

SUMMARY OF THE INVENTION

The present invention identifies the global changes in gene expression associated with liver cancer by examining gene expression in tissue from normal liver, metastatic malignant liver and hepatocellular carcinoma. The present invention also identifies expression profiles which serve as useful diagnostic markers as well as markers that can be used to monitor disease states, disease progression, drug toxicity, drug efficacy and drug metabolism.

The invention includes methods of diagnosing the presence or absence of liver cancer in a patient comprising the step of detecting the level of expression in a tissue sample of two or more genes from Tables 3-9; wherein differential expression of the genes in Tables 3-9 is indicative of liver cancer. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3-5. In some preferred embodiments, the method may include detecting the expression level of one or more genes selected from a group consisting Tetraspan NET-6 protein; collagen, type V, alpha; and glypican 3.

The invention also includes methods of detecting the progression of liver cancer and/or differentiating nonmetastatic from metastatic disease. For instance, methods of the invention include detecting the progression of liver cancer in a patient comprising the step of detecting the level of expression in a tissue sample of two or more genes from 3-9; wherein differential expression of the genes in Tables 3-9 is indicative of liver cancer progression. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3-5.

In some aspects, the present invention provides a method of monitoring the treatment of a patient with liver cancer, comprising administering a pharmaceutical composition to the patient and preparing a gene expression profile from a cell or tissue sample from the patient and comparing the patient gene expression profile to a gene expression from a cell population comprising normal liver cells or to a gene expression profile from a cell population comprising liver cancer cells or to both. In some preferred embodiments, the gene profile will include the expression level of one or more genes in Tables 3-9. In other preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3-5.

In another aspect, the present invention provides a method of treating a patient with liver cancer, comprising administering to the patient a pharmaceutical composition, wherein the composition alters the expression of at least one gene in Tables 3-9, preparing a gene expression profile from a cell or tissue sample from the patient comprising tumor cells and comparing the patient expression profile to a gene expression profile from an untreated cell population comprising liver cancer cells. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3-5.

In one aspect, the present invention provides a method of diagnosing hepatocellular carcinoma in a patient, comprising detecting the level of expression in a tissue sample of two or more genes from Tables 3-9, wherein differential expression of the genes in Tables 3-9 is indicative of hepatocellular carcinoma. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3 or 5

In another aspect, the present invention provides a method of detecting the progression of hepatocellular carcinoma in a patient, comprising detecting the level of expression in a tissue sample of two or more genes from Tables 3-9; wherein differential expression of the genes in Tables 3-9 is indicative of hepatocellular carcinoma progression. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3 or 5.

The present invention also provides materials and methods for monitoring the treatment of a patient with a hepatocellular caricnoma. The present invention provides a method of monitoring the treatment of a patient with hepatocellular carcinoma, comprising administering a pharmaceutical composition to the patient, preparing a gene expression profile from a cell or tissue sample from the patient and comparing the patient gene expression profile to a gene expression from a cell population comprising normal liver cells or to a gene expression profile from a cell population comprising hepatocellular carcinoma cells or to both. In some preferred embodiments, the method may include detecting the level of expression of one or more genes from the genes listed in Tables 3-9. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3 or 5.

In a related aspect, the present invention provides a method of treating a patient with hepatocellular carcinoma, comprising administering to the patient a pharmaceutical composition, wherein the composition alters the expression of at least one gene in Tables 3-

9, preparing a gene expression profile from a cell or tissue sample from the patient comprising hepatocellular carcinoma cells and comparing the patient expression profile to a gene expression profile from an untreated cell population comprising hepatocellular carcinoma cells. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3 or 5.

The present invention provides a method of diagnosing a metastatic liver tumor in a patient, comprising detecting the level of expression in a tissue sample of two or more genes from Tables 3-9, wherein differential expression of the genes in Tables 3-9 is indicative of hepatocellular carcinoma. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 4 or 5.

The present invention provides a method of detecting the progression of a metastatic liver tumor in a patient, comprising detecting the level of expression in a tissue sample of two or more genes from Tables 3-9, wherein differential expression of the genes in Tables 3-9 is indicative of a metastatic liver tumor progression. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 4 or 5.

In a related aspect, the present invention provides a method of monitoring the treatment of a patient with a metastatic liver tumor, comprising administering a pharmaceutical composition to the patient, preparing a gene expression profile from a cell or tissue sample from the patient and comparing the patient gene expression profile to a gene expression from a cell population comprising normal liver cells or to a gene expression profile from a cell population comprising metastatic liver tumor cells or to both. In some preferred embodiments, the method of the present invention may include detecting the expression level of one or more genes selected from the genes listed in Tables 3-9. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 4 or 5.

In some preferred embodiments, the present invention provides a method of treating a patient with a metastatic liver tumor, comprising administering to the patient a pharmaceutical composition, wherein the composition alters the expression of at least one gene in Tables 3-9, preparing a gene expression profile from a cell or tissue sample from the patient comprising metastatic liver tumor cells and comparing the patient expression profile to a gene expression profile from an untreated cell population comprising metastatic liver

tumor cells. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 4 or 5.

The invention also includes methods of differentiating metastatic liver cancer from hepatocellular carcinoma in a patient comprising the step of detecting the level of expression in a tissue sample of two or more genes from Tables 3-9; wherein differential expression of the genes in Tables 3-9 is indicative of metastatic liver cancer rather than hepatocellular carcinoma.

The invention further includes methods of screening for an agent capable of modulating the onset or progression of liver cancer, comprising the steps of exposing a cell to the agent; and detecting the expression level of two or more genes from Tables 3-9. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3-5.

Any of the methods of the invention described above may include the detection of at least 2 genes from the tables. Preferred methods may detect all or nearly all of the genes in the tables. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3-5.

The invention further includes compositions comprising at least two oligonucleotides, wherein each of the oligonucleotides comprises a sequence that specifically hybridizes to a gene in Tables 3-9 as well as solid supports comprising at least two probes, wherein each of the probes comprises a sequence that specifically hybridizes to a gene in Tables 3-9. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3-5.

The invention further includes computer systems comprising a database containing information identifying the expression level in liver tissue of a set of genes comprising at least two genes in Tables 3-9; and a user interface to view the information. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3-5. The database may further include sequence information for the genes, information identifying the expression level for the set of genes in normal liver tissue and malignant tissue (metastatic and nonmetastatic) and may contain links to external databases such as GenBank.

The invention further comprises kits useful for the practice of one or more of the methods of the invention. In some preferred embodiments, a kit may contain one or more

solid supports having attached thereto one or more oligonucleotides. The solid support may be a high-density oligonucleotide array. Kits may further comprise one or more reagents for use with the arrays, one or more signal detection and/or array-processing instruments, one or more gene expression databases and one or more analysis and database management software packages.

Lastly, the invention includes methods of using the databases, such as methods of using the disclosed computer systems to present information identifying the expression level in a tissue or cell of at least one gene in Tables 3-9, comprising the step of comparing the expression level of at least one gene in Tables 3-9 in the tissue or cell to the level of expression of the gene in the database. In some preferred embodiments, one or more genes may be selected from a group consisting of the genes listed in Tables 3-5.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a flow chart showing a schematic representation of the experimental protocol.

Figures 2A-2C are graphs of the number of genes present in all samples as a function of the number of samples for the second sample set. Figure 2A is the Gene Signature Curve for normal liver tissue. Figure 2B is the Gene Signature Curve for metastatic liver tumor samples. Figure 2C is the Gene Signature Curve for hepatocellular carinoma samples.

DETAILED DESCRIPTION

Many biological functions are accomplished by altering the expression of various genes through transcriptional (*e.g.*, through control of initiation, provision of RNA precursors, RNA processing, etc.) and/or translational control. For example, fundamental biological processes such as cell cycle, cell differentiation and cell death, are often characterized by the variations in the expression levels of groups of genes.

Changes in gene expression also are associated with pathogenesis. For example, the lack of sufficient expression of functional tumor suppressor genes and/or the over expression of oncogene/protooncogenes could lead to tumorgenesis or hyperplastic growth of cells (Marshall, (1991) Cell, 64, 313-326; Weinberg, (1991) Science, 254, 1138-1146). Thus, changes in the expression levels of particular genes (*e.g.*, oncogenes or tumor suppressors) serve as signposts for the presence and progression of various diseases.

Monitoring changes in gene expression may also provide certain advantages during drug screening development. Often drugs are screened and prescreened for the ability to interact with a major target without regard to other effects the drugs have on cells. Often such other effects cause toxicity in the whole animal, which prevent the development and use of the potential drug.

The present inventors have examined tissue samples from normal liver, metastatic malignant liver and hepatocellular carcinoma to identify the global changes in gene expression associated with liver cancer. The protocol used is schematically represented in Figure 1. These global changes in gene expression, also referred to as expression profiles, provide useful markers for diagnostic uses as well as markers that can be used to monitor disease states, disease progression, drug toxicity, drug efficacy and drug metabolism.

The present invention provides compositions and methods to detect the level of expression of genes that may be differentially expressed dependent upon the state of the cell, *i.e.*, normal versus cancerous. As used herein, the phrase "detecting the level expression" includes methods that quantitate expression levels as well as methods that determine whether a gene of interest is expressed at all. Thus, an assay which provides a yes or no result without necessarily providing quantification of an amount of expression is an assay that requires "detecting the level of expression" as that phrase is used herein.

Assay Formats

The genes identified as being differentially expressed in liver cancer may be used in a variety of nucleic acid detection assays to detect or quantititate the expression level of a gene or multiple genes in a given sample. For example, traditional Northern blotting, nuclease protection, RT-PCR and differential display methods may be used for detecting gene expression levels. Those methods are useful for some embodiments of the invention. However, methods and assays of the invention are most efficiently designed with array or chip hybridization-based methods for detecting the expression of a large number of genes.

Any hybridization assay format may be used, including solution-based and solid support-based assay formats. Solid supports containing oligonucleotide probes for differentially expressed genes of the invention can be filters, polyvinyl chloride dishes, silicon or glass based chips, etc. Such wafers and hybridization methods are widely available, for example, those disclosed by Beattie (WO 95/11755). Any solid surface to

which oligonucleotides can be bound, either directly or indirectly, either covalently or non-covalently, can be used. A preferred solid support is a high density array or DNA chip. These contain a particular oligonucleotide probe in a predetermined location on the array. Each predetermined location may contain more than one molecule of the probe, but each molecule within the predetermined location has an identical sequence. Such predetermined locations are termed features. There may be, for example, about 2, 10, 100, 1000 to 10,000; 100,000 or 400,000 of such features on a single solid support. The solid support, or the area within which the probes are attached may be on the order of a square centimeter.

Oligonucleotide probe arrays for expression monitoring can be made and used according to any techniques known in the art (see for example, Lockhart *et al.*, (1996) Nat. Biotechnol. 14, 1675-1680; McGall *et al.*, (1996) Proc. Nat. Acad. Sci. USA 93, 13555-13460). Such probe arrays may contain at least two or more oligonucleotides that are complementary to or hybridize to two or more of the genes described herein. Such arrays may also contain oligonucleotides that are complementary or hybridize to at least about 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 50, 70, 100 or or more the genes described herein.

The genes which are assayed according to the present invention are typically in the form of mRNA or reverse transcribed mRNA. The genes may be cloned or not and the genes may be amplified or not. The cloning itself does not appear to bias the representation of genes within a population. However, it may be preferable to use polyA+ RNA as a source, as it can be used with less processing steps.

The sequences of the expression marker genes are in the public databases. Tables 3-9 provide the GenBank accession number for the genes and ESTs identified called either Accession # (Tables 3, 4, and 5) or Fragment Name (Tables 6-9). The sequences of the genes in GenBank are expressly incorporated by reference as are equivalent and related sequences present in GenBank or other public databases. The column labeled "SEQ ID" refers to the sequence identification number correlating the listed gene or EST to its sequence information as provided within the sequence listing of this application.

Probes based on the sequences of the genes described herein may be prepared by any commonly available method. Oligonucleotide probes for assaying the tissue or cell sample are preferably of sufficient length to specifically hybridize only to appropriate, complementary genes or transcripts. Typically the oligonucleotide probes will be at least 10,

12, 14, 16, 18, 20 or 25 nucleotides in length. In some cases longer probes of at least 30, 40, or 50 nucleotides will be desirable.

As used herein, oligonucleotide sequences that are complementary to one or more of the genes described herein, refers to oligonucleotides that are capable of hybridizing under stringent conditions to at least part of the nucleotide sequence of said genes. Such hybridizable oligonucleotides will typically exhibit at least about 75% sequence identity at the nucleotide level to said genes, preferably about 80% or 85% sequence identity or more preferably about 90% or 95% or more sequence identity to said genes.

"Bind(s) substantially" refers to complementary hybridization between a probe nucleic acid and a target nucleic acid and embraces minor mismatches that can be accommodated by reducing the stringency of the hybridization media to achieve the desired detection of the target polynucleotide sequence.

The terms "background" or "background signal intensity" refer to hybridization signals resulting from non-specific binding, or other interactions, between the labeled target nucleic acids and components of the oligonucleotide array (*e.g.*, the oligonucleotide probes, control probes, the array substrate, etc.). Background signals may also be produced by intrinsic fluorescence of the array components themselves. A single background signal can be calculated for the entire array, or a different background signal may be calculated for each target nucleic acid. In a preferred embodiment, background is calculated as the average hybridization signal intensity for the lowest 5% to 10% of the probes in the array, or, where a different background signal is calculated for each target gene, for the lowest 5% to 10% of the probes for each gene. Of course, one of skill in the art will appreciate that where the probes to a particular gene hybridize well and thus appear to be specifically binding to a target sequence, they should not be used in a background signal calculation. Alternatively, background may be calculated as the average hybridization signal intensity produced by hybridization to probes that are not complementary to any sequence found in the sample (*e.g.*, probes directed to nucleic acids of the opposite sense or to genes not found in the sample such as bacterial genes where the sample is mammalian nucleic acids). Background can also be calculated as the average signal intensity produced by regions of the array that lack any probes at all.

The phrase "hybridizing specifically to" refers to the binding, duplexing or hybridizing of a molecule substantially to or only to a particular nucleotide sequence or

sequences under stringent conditions when that sequence is present in a complex mixture (*e.g.*, total cellular) DNA or RNA.

Assays and methods of the invention may utilize available formats to simultaneously screen at least about 100, preferably about 1000, more preferably about 10,000 and most preferably about 1,000,000 or more different nucleic acid hybridizations.

The term "mismatch control" or "mismatch probe" refer to a probe whose sequence is deliberately selected not to be perfectly complementary to a particular target sequence. For each mismatch (MM) control in a high-density array there typically exists a corresponding perfect match (PM) probe that is perfectly complementary to the same particular target sequence. The mismatch may comprise one or more bases.

While the mismatch(s) may be located anywhere in the mismatch probe, terminal mismatches are less desirable as a terminal mismatch is less likely to prevent hybridization of the target sequence. In a particularly preferred embodiment, the mismatch is located at or near the center of the probe such that the mismatch is most likely to destabilize the duplex with the target sequence under the test hybridization conditions.

The term "perfect match probe" refers to a probe that has a sequence that is perfectly complementary to a particular target sequence. The test probe is typically perfectly complementary to a portion (subsequence) of the target sequence. The perfect match (PM) probe can be a "test probe", a "normalization control" probe, an expression level control probe and the like. A perfect match control or perfect match probe is, however, distinguished from a "mismatch control" or "mismatch probe."

As used herein a "probe" is defined as a nucleic acid, capable of binding to a target nucleic acid of complementary sequence through one or more types of chemical bonds, usually through complementary base pairing, usually through hydrogen bond formation. As used herein, a probe may include natural (*i.e.*, A, G, U, C or T) or modified bases (7-deazaguanosine, inosine, etc.). In addition, the bases in probes may be joined by a linkage other than a phosphodiester bond, so long as it does not interfere with hybridization. Thus, probes may be peptide nucleic acids in which the constituent bases are joined by peptide bonds rather than phosphodiester linkages.

The term "stringent conditions" refers to conditions under which a probe will hybridize to its target subsequence, but with only insubstantial hybridization to other sequences or to other sequences such that the difference may be identified. Stringent

conditions are sequence-dependent and will be different in different circumstances. Longer sequences hybridize specifically at higher temperatures. Generally, stringent conditions are selected to be about 5°C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH.

Typically, stringent conditions will be those in which the salt concentration is at least about 0.01 to 1.0 M sodium ion concentration (or other salts) at pH 7.0 to 8.3 and the temperature is at least about 30°C for short probes (*e.g.*, 10 to 50 nucleotide). Stringent conditions may also be achieved with the addition of destabilizing agents such as formamide.

The "percentage of sequence identity" or "sequence identity" is determined by comparing two optimally aligned sequences or subsequences over a comparison window or span, wherein the portion of the polynucleotide sequence in the comparison window may optionally comprise additions or deletions (*i.e.*, gaps) as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. The percentage is calculated by determining the number of positions at which the identical monomer unit (*e.g.*, nucleic acid base or amino acid residue) occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison and multiplying the result by 100 to yield the percentage of sequence identity. Percentage sequence identity when calculated using the programs GAP or BESTFIT (see below) is calculated using default gap weights.

Homology or identity may be determined by **BLAST** (**B**asic **L**ocal **A**lignment **S**earch **T**ool) analysis using the algorithm employed by the programs **blastp**, **blastn**, **blastx**, **tblastn** and **tblastx** (Karlin *et al.*, (1990) Proc. Natl. Acad. Sci. USA 87, 2264-2268 and Altschul, (1993) J. Mol. Evol. 36, 290-300, fully incorporated by reference) which are tailored for sequence similarity searching. The approach used by the **BLAST** program is to first consider similar segments between a query sequence and a database sequence, then to evaluate the statistical significance of all matches that are identified and finally to summarize only those matches which satisfy a preselected threshold of significance. For a discussion of basic issues in similarity searching of sequence databases, see Altschul *et al.*, (1994) Nature Genet. 6, 119-129) which is fully incorporated by reference. The search parameters for **histogram**, **descriptions**, **alignments**, **expect** (*i.e.*, the statistical

significance threshold for reporting matches against database sequences), **cutoff**, **matrix** and **filter** are at the default settings. The default scoring matrix used by **blastp**, **blastx**, **tblastn**, and **tblastx** is the **BLOSUM62** matrix (Henikoff *et al.*, (1992) Proc. Natl. Acad. Sci. USA 89, 10915-10919, fully incorporated by reference). Four **blastn** parameters were adjusted as follows: Q=10 (gap creation penalty); R=10 (gap extension penalty); wink=1 (generates word hits at every winkth position along the query); and gapw=16 (sets the window width within which gapped alignments are generated). The equivalent **Blastp** parameter settings were Q=9; R=2; wink=1; and gapw=32. A **Bestfit** comparison between sequences, available in the GCG package version 10.0, uses DNA parameters GAP=50 (gap creation penalty) and LEN=3 (gap extension penalty) and the equivalent settings in protein comparisons are GAP=8 and LEN=2.

Probe design

One of skill in the art will appreciate that an enormous number of array designs are suitable for the practice of this invention. The high density array will typically include a number of probes that specifically hybridize to the sequences of interest. See WO 99/32660 for methods of producing probes for a given gene or genes. In addition, in a preferred embodiment, the array will include one or more control probes.

High density array chips of the invention include "test probes." Test probes may be oligonucleotides that range from about 5 to about 500 or about 5 to about 50 nucleotides, more preferably from about 10 to about 40 nucleotides and most preferably from about 15 to about 40 nucleotides in length. In other particularly preferred embodiments the probes are about 20 to 25 nucleotides in length. In another preferred embodiment, test probes are double or single strand DNA sequences. DNA sequences are isolated or cloned from natural sources or amplified from natural sources using natural nucleic acid as templates. These probes have sequences complementary to particular subsequences of the genes whose expression they are designed to detect. Thus, the test probes are capable of specifically hybridizing to the target nucleic acid they are to detect.

In addition to test probes that bind the target nucleic acid(s) of interest, the high density array can contain a number of control probes. The control probes fall into three categories referred to herein as (1) normalization controls; (2) expression level controls; and (3) mismatch controls.

Normalization controls are oligonucleotide or other nucleic acid probes that are complementary to labeled reference oligonucleotides or other nucleic acid sequences that are added to the nucleic acid sample. The signals obtained from the normalization controls after hybridization provide a control for variations in hybridization conditions, label intensity, "reading" efficiency and other factors that may cause the signal of a perfect hybridization to vary between arrays. In a preferred embodiment, signals (*e.g.*, fluorescence intensity) read from all other probes in the array are divided by the signal (*e.g.*, fluorescence intensity) from the control probes thereby normalizing the measurements.

Virtually any probe may serve as a normalization control. However, it is recognized that hybridization efficiency varies with base composition and probe length. Preferred normalization probes are selected to reflect the average length of the other probes present in the array, however, they can be selected to cover a range of lengths. The normalization control(s) can also be selected to reflect the (average) base composition of the other probes in the array, however in a preferred embodiment, only one or a few probes are used and they are selected such that they hybridize well (*i.e.*, no secondary structure) and do not match any target-specific probes.

Expression level controls are probes that hybridize specifically with constitutively expressed genes in the biological sample. Virtually any constitutively expressed gene provides a suitable target for expression level controls. Typical expression level control probes have sequences complementary to subsequences of constitutively expressed "housekeeping genes" including, but not limited to the β-actin gene, the transferrin receptor gene, the GAPDH gene, and the like.

Mismatch controls may also be provided for the probes to the target genes, for expression level controls or for normalization controls. Mismatch controls are oligonucleotide probes or other nucleic acid probes identical to their corresponding test or control probes except for the presence of one or more mismatched bases. A mismatched base is a base selected so that it is not complementary to the corresponding base in the target sequence to which the probe would otherwise specifically hybridize. One or more mismatches are selected such that under appropriate hybridization conditions (*e.g.*, stringent conditions) the test or control probe would be expected to hybridize with its target sequence, but the mismatch probe would not hybridize (or would hybridize to a significantly lesser extent). Preferred mismatch probes contain a central mismatch. Thus, for example, where a

probe is a twenty-mer, a corresponding mismatch probe will have the identical sequence except for a single base mismatch (*e.g.*, substituting a G, a C or a T for an A) at any of positions 6 through 14 (the central mismatch).

Mismatch probes thus provide a control for non-specific binding or cross hybridization to a nucleic acid in the sample other than the target to which the probe is directed. Mismatch probes also indicate whether a hybridization is specific or not. For example, if the target is present the perfect match probes should be consistently brighter than the mismatch probes. In addition, if all central mismatches are present, the mismatch probes can be used to detect a mutation. The difference in intensity between the perfect match and the mismatch probe ($I_{(PM)} - I_{(MM)}$) provides a good measure of the concentration of the hybridized material.

Nucleic Acid Samples

As is apparent to one of ordinary skill in the art, nucleic acid samples used in the methods and assays of the invention may be prepared by any available method or process. Methods of isolating total mRNA are also well known to those of skill in the art. For example, methods of isolation and purification of nucleic acids are described in detail in Chapter 3 of Laboratory Techniques in Biochemistry and Molecular Biology: Hybridization With Nucleic Acid Probes, Part I Theory and Nucleic Acid Preparation, Tijssen, (1993) (editor) Elsevier Press. Such samples include RNA samples, but also include cDNA synthesized from a mRNA sample isolated from a cell or tissue of interest. Such samples also include DNA amplified from the cDNA, and an RNA transcribed from the amplified DNA. One of skill in the art would appreciate that it is desirable to inhibit or destroy RNase present in homogenates before homogenates can be used.

Biological samples may be of any biological tissue or fluid or cells from any organism as well as cells raised *in vitro*, such as cell lines and tissue culture cells. Frequently the sample will be a "clinical sample" which is a sample derived from a patient. Typical clinical samples include, but are not limited to, sputum, blood, blood-cells (*e.g.*, white cells), tissue or fine needle biopsy samples, urine, peritoneal fluid, and pleural fluid, or cells therefrom.

Biological samples may also include sections of tissues, such as frozen sections or formalin fixed sections taken for histological purposes.

Forming High Density Arrays.

Methods of forming high density arrays of oligonucleotides with a minimal number of synthetic steps are known. The oligonucleotide analogue array can be synthesized on a solid substrate by a variety of methods, including, but not limited to, light-directed chemical coupling, and mechanically directed coupling (see Pirrung *et al.*, (1992) U.S. Patent No. 5,143, 854; Fodor *et al.*, (1998) U.S. Patent No. 5,800,992; Chee *et al.*, (1998) 5,837,832

In brief, the light-directed combinatorial synthesis of oligonucleotide arrays on a glass surface proceeds using automated phosphoramidite chemistry and chip masking techniques. In one specific implementation, a glass surface is derivatized with a silane reagent containing a functional group, *e.g.*, a hydroxyl or amine group blocked by a photolabile protecting group. Photolysis through a photolithogaphic mask is used selectively to expose functional groups which are then ready to react with incoming 5' photoprotected nucleoside phosphoramidites. The phosphoramidites react only with those sites which are illuminated (and thus exposed by removal of the photolabile blocking group). Thus, the phosphoramidites only add to those areas selectively exposed from the preceding step. These steps are repeated until the desired array of sequences have been synthesized on the solid surface. Combinatorial synthesis of different oligonucleotide analogues at different locations on the array is determined by the pattern of illumination during synthesis and the order of addition of coupling reagents.

In addition to the foregoing, additional methods which can be used to generate an array of oligonucleotides on a single substrate are described in Fodor *et al.*, (1993). WO 93/09668. High density nucleic acid arrays can also be fabricated by depositing premade or natural nucleic acids in predetermined positions. Synthesized or natural nucleic acids are deposited on specific locations of a substrate by light directed targeting and oligonucleotide directed targeting. Another embodiment uses a dispenser that moves from region to region to deposit nucleic acids in specific spots.

Hybridization

Nucleic acid hybridization simply involves contacting a probe and target nucleic acid under conditions where the probe and its complementary target can form stable hybrid duplexes through complementary base pairing (see Lockhart *et al.*, (1999) WO 99/32660).

The nucleic acids that do not form hybrid duplexes are then washed away leaving the hybridized nucleic acids to be detected, typically through detection of an attached detectable label. It is generally recognized that nucleic acids are denatured by increasing the temperature or decreasing the salt concentration of the buffer containing the nucleic acids.

Under low stringency conditions (*e.g.*, low temperature and/or high salt) hybrid duplexes (*e.g.*, DNA-DNA, RNA-RNA or RNA-DNA) will form even where the annealed sequences are not perfectly complementary.

Thus specificity of hybridization is reduced at lower stringency. Conversely, at higher stringency (*e.g.*, higher temperature or lower salt) successful hybridization requires fewer mismatches. One of skill in the art will appreciate that hybridization conditions may be selected to provide any degree of stringency. In a preferred embodiment, hybridization is performed at low stringency, in this case in 6× SSPE-T at 37°C (0.005% Triton x-100) to ensure hybridization and then subsequent washes are performed at higher stringency (*e.g.*, 1× SSPE-T at 37°C) to eliminate mismatched hybrid duplexes. Successive washes may be performed at increasingly higher stringency (*e.g.*, down to as low as 0.25× SSPET at 37°C to 50°C) until a desired level of hybridization specificity is obtained. Stringency can also be increased by addition of agents such as formamide. Hybridization specificity may be evaluated by comparison of hybridization to the test probes with hybridization to the various controls that can be present (*e.g.*, expression level control, normalization control, mismatch controls, etc.).

In general, there is a tradeoff between hybridization specificity (stringency) and signal intensity. Thus, in a preferred embodiment, the wash is performed at the highest stringency that produces consistent results and that provides a signal intensity greater than approximately 10% of the background intensity. Thus, in a preferred embodiment, the hybridized array may be washed at successively higher stringency solutions and read between each wash. Analysis of the data sets thus produced will reveal a wash stringency above which the hybridization pattern is not appreciably altered and which provides adequate signal for the particular oligonucleotide probes of interest.

Signal Detection

The hybridized nucleic acids are typically detected by detecting one or more labels attached to the sample nucleic acids. The labels may be incorporated by any of a number of means well known to those of skill in the art (see Lockhart *et al.*, (1999) WO 99/32660).

Databases

The present invention includes relational databases containing sequence information, for instance for the genes of Tables 3-9, as well as gene expression information in various liver tissue samples. Databases may also contain information associated with a given sequence or tissue sample such as descriptive information about the gene associated with the sequence information, or descriptive information concerning the clinical status of the tissue sample, or the patient from which the sample was derived. The database may be designed to include different parts, for instance a sequences database and a gene expression database. Methods for the configuration and construction of such databases are widely available, for instance, see Akerblom *et al.*, (1999) U.S. Patent 5,953,727, which is herein incorporated by reference in its entirety.

The databases of the invention may be linked to an outside or external database. In a preferred embodiment, as described in Tables 3-9, the external database is GenBank and the associated databases maintained by the National Center for Biotechnology Information (NCBI).

Any appropriate computer platform may be used to perform the necessary comparisons between sequence information, gene expression information and any other information in the database or provided as an input. For example, a large number of computer workstations are available from a variety of manufacturers, such has those available from Silicon Graphics. Client-server environments, database servers and networks are also widely available and appropriate platforms for the databases of the invention.

The databases of the invention may be used to produce, among other things, electronic Northerns to allow the user to determine the cell type or tissue in which a given gene is expressed and to allow determination of the abundance or expression level of a given gene in a particular tissue or cell.

The databases of the invention may also be used to present information identifying the expression level in a tissue or cell of a set of genes comprising at least one gene in Tables 3-9 comprising the step of comparing the expression level of at least one gene in

Tables 3-9 in the tissue to the level of expression of the gene in the database. Such methods may be used to predict the physiological state of a given tissue by comparing the level of expression of a gene or genes in Tables 3-9 from a sample to the expression levels found in tissue from normal liver, malignant liver or hepatocellular carcinoma. Such methods may also be used in the drug or agent screening assays as described below.

Kits

The invention further includes kits combining, in different combinations, high-density oligonucleotide arrays, reagents for use with the arrays, signal detection and array-processing instruments, gene expression databases and analysis and database management software described above. The kits may be used, for example, to predict or model the toxic response of a test compound, to monitor the progression of liver disease states, to identify genes that show promise as new drug targets and to screen known and newly designed drugs as discussed above.

The databases packaged with the kits are a compilation of expression patterns from human or laboratory animal genes and gene fragments (corresponding to the genes of Table 3-9). Data is collected from a repository of both normal and diseased animal tissues and provides reproducible, quantitative results, *i.e.*, the degree to which a gene is up-regulated or down-regulated under a given condition.

The kits may used in the pharmaceutical industry, where the need for early drug testing is strong due to the high costs associated with drug development, but where bioinformatics, in particular gene expression informatics, is still lacking. These kits will reduce the costs, time and risks associated with traditional new drug screening using cell cultures and laboratory animals. The results of large-scale drug screening of pre-grouped patient populations, pharmacogenomics testing, can also be applied to select drugs with greater efficacy and fewer side-effects. The kits may also be used by smaller biotechnology companies and research institutes who do not have the facilities for performing such large-scale testing themselves.

Databases and software designed for use with use with microarrays is discussed in Balaban *et al.*, U.S. Patent Nos. 6,229,911, a computer-implemented method for managing information, stored as indexed tables, collected from small or large numbers of microarrays, and 6,185,561, a computer-based method with data mining capability for collecting gene

expression level data, adding additional attributes and reformatting the data to produce answers to various queries. Chee *et al.*, U.S. Patent No. 5,974,164, disclose a software-based method for identifying mutations in a nucleic acid sequence based on differences in probe fluorescence intensities between wild type and mutant sequences that hybridize to reference sequences.

Diagnostic Uses for the Liver Cancer Markers

As described above, the genes and gene expression information provided in Tables 3-9 may be used as diagnostic markers for the prediction or identification of the malignant state of the liver tissue. For instance, a liver tissue sample or other sample from a patient may be assayed by any of the methods described above, and the expression levels from a gene or genes from the Tables, in particular the genes in Tables 3-5, may be compared to the expression levels found in normal liver tissue, tissue from metastatic liver cancer or hepatocellular carcinoma tissue. Expression profiles generated from the tissue or other sample that substantially resemble an expression profile from normal or diseased liver tissue may be used, for instance, to aid in disease diagnosis. Comparison of the expression data, as well as available sequence or other information may be done by researcher or diagnostician or may be done with the aid of a computer and databases as described above.

Use of the Liver Cancer Markers for Monitoring Disease Progression

As described above, the genes and gene expression information provided in Tables 3-9 may also be used as markers for the monitoring of disease progression, for instance, the development of liver cancer. For instance, a liver tissue sample or other sample from a patient may be assayed by any of the methods described above, and the expression levels in the sample from a gene or genes from or 3-9 may be compared to the expression levels found in normal liver tissue, tissue from metastatic liver cancer or hepatocellular carcinoma tissue. Comparison of the expression data, as well as available sequence or other information may be done by researcher or diagnostician or may be done with the aid of a computer and databases as described above.

Use of the Liver Cancer Markers for Drug Screening

According to the present invention, the genes identified in Tables 3-9 may be used as markers to evaluate the effects of a candidate drug or agent on a cell, particularly a cell undergoing malignant transformation, for instance, a liver cancer cell or tissue sample. A candidate drug or agent can be screened for the ability to simulate the transcription or expression of a given marker or markers (drug targets) or to down-regulate or counteract the transcription or expression of a marker or markers. According to the present invention, one can also compare the specificity of drugs' effects by looking at the number of markers which the drugs have and comparing them. More specific drugs will have fewer transcriptional targets. Similar sets of markers identified for two drugs indicates a similarity of effects.

Assays to monitor the expression of a marker or markers as defined in Tables 3-9 may utilize any available means of monitoring for changes in the expression level of the nucleic acids of the invention. As used herein, an agent is said to modulate the expression of a nucleic acid of the invention if it is capable of up- or down-regulating expression of the nucleic acid in a cell.

In one assay format, gene chips containing probes to at least two genes from Tables 3-9 may be used to directly monitor or detect changes in gene expression in the treated or exposed cell as described in more detail above. In another format, cell lines that contain reporter gene fusions between the open reading frame and/or the 3' or 5' regulatory regions of a gene in Tables 3-9 and any assayable fusion partner may be prepared. Numerous assayable fusion partners are known and readily available including the firefly luciferase gene and the gene encoding chloramphenicol acetyltransferase (Alam *et al.*, (1990) Anal. Biochem. 188, 245-254). Cell lines containing the reporter gene fusions are then exposed to the agent to be tested under appropriate conditions and time. Differential expression of the reporter gene between samples exposed to the agent and control samples identifies agents which modulate the expression of the nucleic acid.

Additional assay formats may be used to monitor the ability of the agent to modulate the expression of a gene identified in Tables 3-9. For instance, as described above, mRNA expression may be monitored directly by hybridization of probes to the nucleic acids of the invention. Cell lines are exposed to the agent to be tested under appropriate conditions and time and total RNA or mRNA is isolated by standard procedures such those disclosed in Sambrook *et al.*, (1989) Molecular Cloning - A Laboratory Manual, Cold Spring Harbor

Laboratory Press).

In another assay format, cells or cell lines are first identified which express the gene products of the invention physiologically. Cell and/or cell lines so identified would be expected to comprise the necessary cellular machinery such that the fidelity of modulation of the transcriptional apparatus is maintained with regard to exogenous contact of agent with appropriate surface transduction mechanisms and/or the cytosolic cascades. Such cell lines may be, but are not required to be, derived from liver tissue. Further, such cells or cell lines may be transduced or transfected with an expression vehicle (*e.g.*, a plasmid or viral vector) construct comprising an operable non-translated 5'-promoter containing end of the structural gene encoding the instant gene products fused to one or more antigenic fragments, which are peculiar to the instant gene products, wherein said fragments are under the transcriptional control of said promoter and are expressed as polypeptides whose molecular weight can be distinguished from the naturally occurring polypeptides or may further comprise an immunologically distinct tag. Such a process is well known in the art (see Sambrook *et al.*, (1989) Molecular Cloning - A Laboratory Manual, Cold Spring Harbor Laboratory Press).

Cells or cell lines transduced or transfected as outlined above are then contacted with agents under appropriate conditions; for example, the agent comprises a pharmaceutically acceptable excipient and is contacted with cells comprised in an aqueous physiological buffer such as phosphate buffered saline (PBS) at physiological pH, Eagles balanced salt solution (BSS) at physiological pH, PBS or BSS comprising serum or conditioned media comprising PBS or BSS and serum incubated at 37°C. Said conditions may be modulated as deemed necessary by one of skill in the art. Subsequent to contacting the cells with the agent, said cells will be disrupted and the polypeptides of the lysate are fractionated such that a polypeptide fraction is pooled and contacted with an antibody to be further processed by immunological assay (*e.g.*, ELISA, immunoprecipitation or Western blot). The pool of proteins isolated from the "agent-contacted" sample will be compared with a control sample where only the excipient is contacted with the cells and an increase or decrease in the immunologically generated signal from the "agent-contacted" sample compared to the control will be used to distinguish the effectiveness of the agent.

Another embodiment of the present invention provides methods for identifying agents that modulate the levels, concentration or at least one activity of a protein(s) encoded

by the genes in Tables 3-9. Such methods or assays may utilize any means of monitoring or detecting the desired activity.

In one format, the relative amounts of a protein of the invention between a cell population that has been exposed to the agent to be tested compared to an un-exposed control cell population may be assayed. In this format, probes such as specific antibodies are used to monitor the differential expression of the protein in the different cell populations. Cell lines or populations are exposed to the agent to be tested under appropriate conditions and time. Cellular lysates may be prepared from the exposed cell line or population and a control, unexposed cell line or population. The cellular lysates are then analyzed with the probe, such as a specific antibody.

Agents that are assayed in the above methods can be randomly selected or rationally selected or designed. As used herein, an agent is said to be randomly selected when the agent is chosen randomly without considering the specific sequences involved in the association of the a protein of the invention alone or with its associated substrates, binding partners, etc. An example of randomly selected agents is the use a chemical library or a peptide combinatorial library, or a growth broth of an organism.

As used herein, an agent is said to be rationally selected or designed when the agent is chosen on a nonrandom basis which takes into account the sequence of the target site and/or its conformation in connection with the agents action. Agents can be rationally selected or rationally designed by utilizing the peptide sequences that make up these sites. For example, a rationally selected peptide agent can be a peptide whose amino acid sequence is identical to or a derivative of any functional consensus site.

The agents of the present invention can be, as examples, peptides, small molecules, vitamin derivatives, as well as carbohydrates. Dominant negative proteins, DNA encoding these proteins, antibodies to these proteins, peptide fragments of these proteins or mimics of these proteins may be introduced into cells to affect function. "Mimic" as used herein refers to the modification of a region or several regions of a peptide molecule to provide a structure chemically different from the parent peptide but topographically and functionally similar to the parent peptide (see Grant, (1995) in Molecular Biology and Biotechnology Meyers (editor) VCH Publishers). A skilled artisan can readily recognize that there is no limit as to the structural nature of the agents of the present invention.

Without further description, it is believed that one of ordinary skill in the art can, using the preceding description and the following illustrative examples, make and utilize the compounds of the present invention and practice the claimed methods. The following working examples therefore, specifically point out the preferred embodiments of the present invention, and are not to be construed as limiting in any way the remainder of the disclosure.

EXAMPLES

Example 1: Tissue Sample Acquisition and Preparation

Figure 1 outlines the experimental protocol used. Liver tissue samples were excised and snap frozen in liquid nitrogen. The clinical data for each of the samples included in this study are outlined in Table 1. The sample set was composed of eight samples of normal liver tissue (N1-N8), five samples of metastatic adenocarcinoma arising from rectum (designated M1 and M3) and colon (M2, M4 and M5) tissues and six samples of primary hepatocellular carcinomas. Samples were named according to type of tissue: HCC=hepatocellular carcinoma, M=metastatic, N=normal. Table 1 include the TNM classification (the American Joint Committee on Cancer's system of classifying cancers) of the tissues used as samples where T refers to the extent of the primary tumor, N refers to the absence or presence and extent of regional lymph node metastasis, and M refers to the absence or presence of distant metastasis. Numbers following T, N, and M refer to the size of the primary tumor and the amount of vascular invasion, where 0=no evidence of tumor, lymph node involvement or metastasis, 4=multiple tumors involved, and x=cannot be assessed. Histopathologic grade (Table 1) is a qualitative assessment of differentiation of a tumor, where G1=most differentiated and G4=undifferentiated. Clinical stage (Table 1) characterizes the anatomic extent of disease in the patient from whom the sample was taken, where I and II are early stages, III and IV are late stages.

With minor modifications, the sample preparation protocol followed the Affymetrix GeneChip Expression Analysis Manual. Frozen tissue was first ground to powder using the Spex Certiprep 6800 Freezer Mill. Total RNA was then extracted using Trizol (Life Technologies). The total RNA yield for each sample (average tissue weight of 300 mg) was 200-500 μg. Next, mRNA was isolated using the Oligotex mRNA Midi kit (Qiagen). Since the mRNA was eluted in a final volume of 400 μl, an ethanol precipitation step was required

to bring the concentration to 1 μg/μl. Using 1-5 μg of mRNA, double stranded cDNA was created using the SuperScript Choice system (Gibco-BRL). First strand cDNA synthesis was primed with a T7-(dT_{24}) oligonucleotide. The cDNA was then phenol-chloroform extracted and ethanol precipitated to a final concentration of 1 μg/μl.

From 2 μg of cDNA, cRNA was synthesized according to standard procedures. To biotin label the cRNA, nucleotides Bio-11-CTP and Bio-16-UTP (Enzo Diagnostics) were added to the reaction. After a 37°C incubation for six hours, the labeled cRNA was cleaned up according to the Rneasy Mini kit protocol (Qiagen). The cRNA was then fragmented (5× fragmentation buffer: 200 mM Tris-Acetate (pH 8.1), 500 mM KOAc, 150 mM MgOAc) for thirty-five minutes at 94°C.

55 μg of fragmented cRNA was hybridized on the human Hu35k set and the HuGeneFL array for twenty-four hours at 60 rpm in a 45°C hybridization oven The chips were washed and stained with Streptavidin Phycoerythrin (SAPE) (Molecular Probes) in Affymetrix fluidics stations. To amplify staining, SAPE solution was added twice with an anti-streptavidin biotinylated antibody (Vector Laboratories) staining step in between. Hybridization to the probe arrays was detected by fluorometric scanning (Hewlett Packard Gene Array Scanner). Following hybridization and scanning, the microarray images were analyzed for quality control, looking for major chip defects or abnormalities in hybridization signal. After all chips passed QC, the data was analyzed using Affymetrix GeneChip software (v3.0), and Experimental Data Mining Tool (EDMT) software (v1.0).

Example 2: Gene Expression Analysis

All samples were prepared as described and hybridized onto the Affymetrix HuGeneFL array and the Human Hu35k set of arrays. Each chip contains 16-20 oligonucleotide probe pairs per gene or cDNA clone. These probe pairs include perfectly matched sets and mismatched sets, both of which are necessary for the calculation of the average difference. The average difference is a measure of the intensity difference for each probe pair, calculated by subtracting the intensity of the mismatch from the intensity of the perfect match. This takes into consideration variability in hybridization among probe pairs and other hybridization artifacts that could affect the fluorescence intensities. Using the average difference value that has been calculated, the GeneChip software then makes an absolute call for each gene or EST.

The absolute call of present, absent or marginal is used to generate a Gene Signature, a tool used to identify those genes that are commonly present or commonly absent in a given sample set, according to the absolute call. For each set of samples, a median average difference was figured using the average differences of each individual sample within the set. The median average difference must be greater than 150 to assure that the expression level is well above the background noise of the hybridization. For the purposes of this study, only the genes and ESTs with a median average difference greater than 150 have been further studied in detail.

The Gene Signature for one set of samples is compared to the Gene Signature of another set of samples to determine the Gene Signature Differential. This comparison identifies the genes that are consistently present in one set of samples and consistently absent in the second set of samples.

The Gene Signature Curve is a graphic view of the number of genes consistently present in a given set of samples as the sample size increases, taking into account the genes commonly expressed among a particular set of samples, and discounting those genes whose expression is variable among those samples. The curve is also indicative of the number of samples necessary to generate an accurate Gene Signature. As the sample number increases, the number of genes common to the sample set decreases. The curve is generated using the positive Gene Signatures of the samples in question, determined by adding one sample at a time to the Gene Signature, beginning with the sample with the smallest number of present genes and adding samples in ascending order. The curve displays the sample size required for the most consistency and the least amount of expression variability from sample to sample. The point where this curve begins to level off represents the minimum number of samples required for the Gene Signature. Graphed on the x-axis is the number of samples in the set, and on the y-axis is the number of genes in the positive Gene Signature.

Example 3: Gene Expression Analysis of Normal Liver Tissue

The gene expression patterns and Gene Signature were individually determined for each sample set: eight samples with normal liver pathology, six samples whose pathology indicated the primary malignancy to be hepatocellular carcinoma, and five samples whose primary colorectal adenocarcinoma had metastasized to the liver. The Gene Signatures obtained for the sample set are shown in Figure 2

The Gene Signature considers the present and absent genes alone, and does not take into consideration those that have been called marginal. Table 2 shows the numbers of present genes, called the positive Gene Signature, and the number of absent genes, called the negative Gene Signature, for each of the three sets of samples.

The Gene Signature is the set of genes that are commonly present or commonly absent in N-1 samples of a given sample set. The positive Gene Signature for the normal liver tissues contains 6,213 genes and ESTs. This same set of normal samples did not show any detectable level of expression of 24,900 genes. Many of the genes and ESTs in this positive Gene Signature are housekeeping genes or structural genes that are not only expressed in the liver, but are ubiquitously expressed in tissues throughout the body. Within this positive Gene Signature are also those genes whose expression is specifically restricted to normal liver tissue and those genes required for the liver to function at its normal capacities. It is the group of genes unique to the liver whose expression levels are most likely to change during tumorigenesis. Whether up-regulated or down-regulated or turned completely on or turned completely off, the changes in expression of these vital genes very likely contributes to the drastic changes in liver function caused by the transformation of normal liver cells into cancerous cells.

Example 4: Gene Expression Analysis of Malignant Liver Tissue

There are 8,479 genes and ESTs in the positive Gene Signature for the HCC tumors, and a total of 23,233 genes and ESTs are included in the negative Gene Signature of the HCC samples. This negative Gene Signature includes all the genes that have been completely turned off during tumorigenesis, as well as those genes that are not usually expressed in liver tissue. These results include a number of genes and ESTs that are not regularly expressed in liver tissues, but through the process of tumor production, their expression patterns have been dramatically altered from no detectable level of expression to some significant level of expression in comparison with the normal liver.

The colorectal metastases in the liver commonly express 5,102 genes and ESTs, and do not show expression of 30,455 additional genes and ESTs. As with the negative Gene Signature for the HCC sample set, the genes included in this data set are generally not expressed in liver tissue, whether tumor or normal tissue. The 5,102 in the sample set of metastatic tumors also identify those genes with expression levels that have been changed

from off to on as a result of tumor formation.

Example 5: Analysis of Gene Expression Profiles

A differential comparison of the genes and ESTs expressed in the normals and the two different types of liver tumors identifies a subset of the genes included in the positive Gene Signatures that are uniquely expressed in each sample set. This Gene Signature Differential highlights genes whose expression profiles have most dramatically changed in the transformation from normal to diseased liver cells. The parameters for these analyses were set to accommodate variation in expression of one eight normal samples and one of the six HCC samples or one of the 5 metastatic tumor samples, such that the genes categorized as unique to normal were called present by the software in seven of eight (87%) normal liver samples and were also called absent in five of six HCC (83%) or four of five (80%) metastatic liver tumor. Conversely, the genes categorized as unique to each set of tumors as compared to the normal livers were called present in five of six HCC (83%) or four of five (80%) metastatic tumor samples and absent in seven of eight normal livers (87%).

The Gene Signature Differential comparing the normal livers to those with metastatic tumors identified a total of 903 sequences expressed only in normal liver tissue. The number of genes or ESTs that meet the median average difference minimum of 150 is 449, of which 289 are genes and the number of ESTs is 160. The remaining ESTs and genes may be indistinguishable from the background noise of the hybridization. The same comparison of normals versus metastatic tumors demonstrates that in the metastatic tumor samples there are 296 uniquely expressed sequences. Those that meet the median average difference minimum requirement are 83 genes and 72 ESTs. Those genes and ESTs expressed in metastatic and not in normal liver tissue are shown in Table 9A and those present in normal liver tissue and not metastatic tissue Table 9B. Numerous genes with differing expression levels in metastatic liver tumor tissue compared to normal tissue were identified. The fifteen genes whose expression level was most different in metastatic as compared to normal tissue are shown in Table 4. Those with the most increased expression are in Table 4A and those with the most decreased expression are in Table 4B. Expression levels were determined by comparing the mean expression values of individual genes in tumor and normal liver samples. Fold change was calculated as a ratio with a p value given as a measure of statistical significance. Fold change is considered significant for a given

gene or EST when it is greater than 3.0 with a p value <0.05. Only the characterized genes have been listed; the ESTs with similar fold changes are not presented here. Asterisk (*) in Table 4 denotes those genes that were also identified in the Gene Signature differential between metastatic liver carcinoma and normal liver tissue. A complete listing of all the genes and ESTs with at least a three-fold change in expression is shown in Table 6. Table 6A contains those genes and ESTs whose expression level increased in metastatic tissue relative to normal tissue and Table 6B contains those genes and ESTs whose expression level decreased.

The Gene Signature Differential between the normal liver samples and the HCC samples identifies a total of 47 unique expressers in the normals, 23 with an median average difference of 150 ,13 of which are named gene and 10 of which are ESTs. When comparing the expression of the HCC samples with the normal livers, there are 243 genes and ESTs only expressed in the HCC samples.

Those genes and ESTs expressed in HCC and not in normal liver tissue are shown in Table 8A and those present in normal liver tissue and not HCC are shown in Table 8B. Numerous genes with differing expression levels in HCC compared to normal tissue were identified. The fifteen genes whose expression level was most different in HCC as compared to normal tissue are shown in Table 3. Those with the most increased expression are in Table 3A and those with the most decreased expression are in Table 3B. Expression levels were determined by comparing the mean expression values of individual genes in tumor and normal liver samples. Fold change was calculated as a ratio with a p value given as a measure of statistical significance. Fold change is considered significant for a given gene or EST when it is greater than 3.0 with a p value <0.05. Only the characterized genes have been listed; the ESTs with similar fold changes are not presented here. Asterisk (*) in Table 3 denotes those genes that were also identified in the Gene Signature differential between hepatocellular carcinoma and normal liver tissue. A complete listing of all the genes and ESTs with at least a three-fold change in expression is shown in Table 7. Table 7A contains those genes and ESTs whose expression level increased in hepatocellular carcinoma tissue relative to normal tissue and Table 7B contains those genes and ESTs whose expression level decreased.

Analysis of sample set identified 24 ESTs and 42 genes that are expressed in both metastatic liver tumors and hepatocellular carcinomas, but not in normal liver tissues. The

fifteen genes with the most increase in expression level in both types of cancer are shown in Table 5. Expression levels were determined by comparing the mean expression values of individual genes in tumor and normal liver samples. The mean expression value for HCC and metastatic carcinomas was greater than 250, and included only those genes that showed a fold change greater than 3 with significant p values for both sets of tumors. No detectable level of expression was found in the normal liver tissues for these genes. Only the characterized genes have been listed; the ESTs with similar fold changes that are unique to the tumors are not presented here.

Differential gene expression patterns between normal liver samples and hepatocellular carcinomas and between normal livers and metastatic liver tumors were examined. Genes uniquely expressed by each of the groups individually were identified, as well as those genes that are commonly expressed among liver tumors, whether primary hepatocellular carcinomas or metastatic liver tumors.

Example 6: Association of Liver Cancer with Specific Gene Expression

The present inventors have closely examined a number of the tumor-expressing genes to determine if their expression patterns correlate with previous reports published in the literature, and to define a logical relationship between the gene and hepatocarcinogenesis. A number of genes that have previously been associated with either liver cancer or other types of cancers were identified, as well as numerous genes that have not been linked to cancers in any previous studies.

842 genes and ESTs that are up-regulated in hepatocellular carcinomas were identified when compared with normal liver tissue. One such gene is PTTG1, pituitary tumor-transforming gene 1, or securin, an oncogene that inhibits sister chromatid separation during anaphase. Normal tissues show little or no PTTG1 expression, but high levels of expression have been associated with various tumors, including liver tumors, and carcinoma cell lines. Overexpression in NIH3T3 cells resulted in transformation, and these cells caused the formation of tumors when injected into mice. The mechanism by which this tumorigenic activity takes place is postulated to be through the missegregation of sister chromatids, resulting in aneuploidy and, therefore, genetic instability. Our data further support this overexpression of PTTG1 in hepatocellular carcinoma, with a fold change of 10.7 (P=0.00052), and no detectable level of expression in normal tissues, as identified by

the differential comparison of the consensus patterns of gene expression of these two sample sets.

Galectin 3, LGALS3, one of a family of beta-galactoside-binding animal lectins, is significantly overexpressed both in primary hepatocellular carcinoma and metastatic liver carcinomas with fold changes of 6.8 (P=0.00103) and 27.1 (P=0.00001), respectively. Expression of LGALS3 has been associated with tumor growth, progression, and metastasis, as well as cell-cell and cell-matrix interactions and inflammatory processes. Although expression studies have revealed no detectable level of galectin-3 in normal liver cells, samples from patients with hepatocellular carcinoma revealed considerable levels of LGALS3 expression. The abnormal expression of this lectin may be an early event in the process of transformation of normal cells to tumor cells, or it may impart an increased capacity for these tumor cells to survive and proliferate. Consistent with the reports in the art, an increased expression level was found in both types of tumor, but higher concentrations of galectin-3 were observed in liver metastates from colorectal tumors than in the primary HCC tumors.

Another gene that is overexpressed in both hepatocellular carcinoma and metastatic colorectal adenocarcinomas with fold changes of 12.2 (P=0.00169) and 58.0 (P=0.00063), respectively, is solute carrier family 2, member 3, or glucose transporter 3 (GLUT3). It is one of a family of transmembrane proteins that function as facilitative glucose transporters, which has a unique specificity for brain and neuronal tissues. Glucose uptake and metabolism are known to be increased in carcinoma cells compared to normal cells. Glucose transporter expression may be elevated in response to the increase in glucose utilization seen in actively proliferating cells, like those of tumors. Conversely, the high levels of glucose transporter expression may be responsible for the enhanced influx of glucose into the tumor cells. Various reports have indicated increased expression of one or more of the family of glucose transporters in malignancies, including those of the brain, esophagus, colon, pancreas, liver, breast, lung, bladder, ovary, testis, skin, head and neck, kidney, and gastric tumors. It has been reported that metastatic liver carcinomas have even higher levels of GLUT3 expression than primary tumors. Consistent with previous studies, the current data confirm the significant overexpression of GLUT3 both in primary liver cancer, hepatocellular carcinoma, and in tumors that have metastasized from the colon and rectum.

One of the significantly underexpressed genes identified by comparing the expression profiles of hepatocellular carcinomas and metastatic liver tumors with that of normal liver tissue is metallothionein 1L. The expression level in HCC is 26.9 fold lower than that of normal (P=0.00999), and in metastatic colorectal adenocarcinomas it is down-regulated 66.5 fold (P=0.00415). Metallothioneins are heavy metal binding proteins that are involved in detoxification of metals, zinc and copper metabolism cellular adaptation mechanisms, and may be involved in regulating apoptosis. Colorectal adenocarcinoma that has metastasized to the liver has been specifically reported to express less metallothionein than normal liver tissue. Comparison of the consensus patterns of gene expression between metastatic liver samples and normal liver samples show no significant level of MT1L expression in the tumors. Furthermore, additional work has determined that human hepatocellular carcinomas contain much lower levels of metallothioneins than normal liver tissue, and that this decrease correlates with the degree of differentiation and concentrations of copper and zinc in the cells. By comparing the expression profiles of hepatocellular carcinoma and normal liver tissue, this significant reduction in MT1L expression in HCC was confirmed.

A number of enzymes belonging to the family of cytochrome P450s are drastically underexpressed in the two sets of liver tumors in comparison with the normal liver tissue. For example, expression of CYP2A6 is decreased in HCC with a fold change of 14.2 (P=0.0307), and in metastatic tumors with a fold change of 69.9 (P=0). CYP8B1 is down-regulated 19.3 fold (P=0.00807) in HCC and 65.1 fold (P=0.0039) in liver metastases. In addition to these commonly down-regulated cytochrome P450s, in HCC samples CYP2B is underexpressed 17.9 fold (P=0.01469), and in the metastatic liver tumors CYP2C9 and CYP2A7 are underexpressed 84.7 fold (P=0.00327) and 72.0 fold (P=0), respectively. Several of these genes are also identified by the differential comparison between expression profiles of tumor and normal, confirming the significant decrease in expression in tumor tissues. Many of these P450 enzymes are critical players in the metabolism of carcinogens, drugs, and other chemical compounds, that are expressed in normal liver.

In addition to genes that are underexpressed in metastatic adenocarcinomas in the liver, more than 1000 genes and ESTs that are overexpressed specifically in these tumors were identified. Two of the most highly up-regulated are claudin 4, also known as clostridium perfringens enterotoxin receptor 1 (fold change 84.4, P=0) and occludin (fold

change 43.1, P=0). Both of these genes are tight junction proteins, responsible for the formation and maintenance of continuous seals around epithelial cells to form a physical barrier that blocks the free passage of water and solutes through the paracellular space. More specifically, claudin-4 is one member of a family of transmembrane proteins that comprise tight junction strands, and occludin is a cell adhesion molecule. Claudins likely function as paracellular channels, regulating the flow of ions and solutes into and out of the paracellular space. Tight junction proteins also contribute to the regulation of the cellular processes of cell growth and differentiation. Permeability of tight junctions has been associated with tumor formation, where a breakdown in the barrier function of tight junctions allows an increase in the cellular permeability. This breakdown then opens the tight junction barrier, permitting invasion by tumor cells. It has been reported that tight junctions of colon tumors leak more than do the tight junctions of normal colon. A complete loss of tight juncton function and a loss of cell-cell contact growth control has been seen in cells that had been transfected with oncogenic Raf-1, and expression levels of occludin and another claudin are lower in these cells. Occludin expression has been up-regulated *in vitro* by the addition of various fatty acids that have anti-cancer effects, decreasing the paracellular permeability. The extreme down-regulation of occludin and claudin-4 in metastatic liver tumors is strongly supported by the reports of tight junction breakdown in tumor tissues.

The present study identified 93 significantly up-regulated genes in both primary HCC and metastatic liver tumors that were not found to have any detectable level of expression in the normal samples. Serine protease inhibitor, Kazal type I (SPINK1), also called pancreatic secretory trypsin inhibitor (PSTI) or tumor-associated trypsin inhibitor (TATI), is one such gene. It is highly expressed in the cells of normal pancreas and in the mucosa of the gastrointestinal tract where it offers protection from proteolytic breakdown. A marked increase in expression is seen in various pancreatic diseases and in tumors of different tissues, including gastric carcinomas, colorectal cancers, and other neoplastic tissues. This increase is presumably due to the elevated expression of trypsin in the tumors, and not related to amplification or rearrangements within the gene. SPINK1 is also considered a valuable marker for a number of solid tumors. An elevation of SPINK1 in the blood of patients with hepatocellular carcinoma has been seen. Furthermore, they suggest that the level of expression correlates with the extent of tumor, such that this heightened

expression level could be indicative of HCC under certain conditions. In keeping with this report of overexpression in these tumors, the present expression data show the levels of expression of this gene in HCC samples to be 28.9 times higher than normal (P=0.00003), and in metastatic liver tumors the expression level is 9.8 times higher than normal (P=0.03697).

Midkine is one of a family of heparin-binding growth factors, inducible by retinoic acid, and is actively involved in cell-cell interactions and angiogenesis. The expression pattern of midkine is highly restricted in normal adult tissues, and no expression has been reported in normal adult liver, although its expression is required during embryogenesis for normal development. However, it is expressed in moderate to high levels in many tumors, including Wilm's tumors of the kidney, stomach, colon, pancreas, lung, esophagus, breast, and liver tumors. The present data confirm these reports, showing a significant overexpression of midkine in hepatocellular carcinoma samples (fold change 9.9, P=0.02104) and in liver metastases (fold change 10.4, P=0.01818), but no noticeable expression in normal liver.

Stathmin, leukemia-associated phosphoprotein 18, is a phosphoprotein whose expression pattern and phosphorylation status are controlled by extracellular signals responsible for the regulation of the processes of cell proliferation and differentiation. It is also involved in the regulation of cell division via the destabilization of microtubules. When comparing expression levels between non-malignant tissues and malignant tissues, the tumors generally show a significant up-regulation of this phosphoprotein, specifically lymphomas, leukemias, breast and prostate tumors. One reason proposed for this elevated expression in cancer cells is the dissimilarity in the rates of cell proliferation and states of differentiation between normal and tumor cells. In both HCC samples and metastatic adenocarcinomas, significant up-regulation of stathmin, 9.4 fold in HCC (P=0.00015) and 4.8 fold in metastatic tumors (P=0.00514) was seen.

Both the genes and ESTs described here will provide valuable information for the identification of new drug targets against liver carcinomas, and that information may be extended for use in the study of carcinogenesis in other tissues. These sequences may be used in the methods of the invention or may be used to produce the probes and arrays of the invention.

Although the present invention has been described in detail with reference to examples above, it is understood that various modifications can be made without departing from the spirit of the invention. Accordingly, the invention is limited only by the following claims. All cited patents, applications and publications referred to in this application are herein incorporated by reference in their entirety.